The Hannan-Quinn Proposition for Linear Regression.

Joe Suzuki*
Department of Mathematics, Osaka University†

MSC2010: 62J05, 62M10

Abstract

We consider the variable selection problem in linear regression. Suppose that we have a set of random variables $X_1, \cdots, X_m, Y, \epsilon$ such that $Y = \sum_{k \in \pi} \alpha_k X_k + \epsilon$ with $\pi \subseteq \{1, \cdots, m\}$ and $\alpha_k \in \mathbb{R}$ unknown, and ϵ is independent of any linear combination of $X_1, \cdots, X_m$. Given actually emitted n examples $\{(x_{i,1} \cdots, x_{i,m}, y_i)\}_{i=1}^n$ emitted from $(X_1, \cdots, X_m, Y)$, we wish to estimate the true π using information criteria in the form of $H + (k/2)d_n$, where H is the likelihood with respect to π multiplied by -1, and $\{d_n\}$ is a positive real sequence. If d_n is too small, we cannot obtain consistency because of overestimation. For autoregression, Hannan-Quinn proved that, in their setting of H and k, the rate $d_n = 2 \log \log n$ is the minimum satisfying strong consistency. This paper solves the statement affirmative for linear regression as well which has a completely different setting.

Keywords

Hannan-Quinn, linear regression, the law of iterated logarithms, strong consistency, information criteria, model selection.

1 Introduction

Information criteria such as AIC, MDL/BIC are used for problems in model selection, and each problem is associated with estimating how many independent parameters exist from given finite examples: on how many variables another variable depends in linear regression (LR); on how many previous variables the subsequent variable depends on in auto regression (AR), etc.

For each model g, we evaluate two factors:

1. How well the examples explain the model g; and

2. How simple the model g is.

*joe@suzuki.email.ne.jp
†Toyonaka, Osaka 560-0043, Japan.

and balance them numerically. Let $\{d_n\}_{n=1}^{\infty}$ be nonnegative reals such that $d_n/n \to 0$, $H(g)$ the empirical entropy which is the maximum likelihood multiplied by (-1), and $k(g)$ the number of parameters in model g. By information criteria, we mean the quantity

$$H(g) + \frac{k(g)}{2} d_n \ , \tag{1}$$

and we estimate the model g by finding one with the minimum value. For example, $d_n = 2$ for AIC, and $d_n = \log n$ for MDL/BIC. Hence, information criteria exist as many as sequences $\{d_n\}_{n=1}^{\infty}$, so it is impossible to list all of information criteria in the form of (1).

In model selection, in particular for theoretical analyses, we often discuss if consistency holds for each $\{d_n\}$, namely, if a sequence of selected models converges to the correct one as $n \to \infty$ in the following senses:

1. the probability of the selected model for each n being correct converges to one (weakly consistent), and
2. the set (event) of infinite sequences in which at most a finite number of errors occur has probability one (strongly consistent).

Although both properties are satisfied in MDL/BIC ($d_n = \log n$), however, none of the two are satisfied in AIC ($d_n = 2$). In general, if d_n is too small, strong consistency is not obtained because of overestimation.

This paper addresses the minimum order of $\{d_n\}$ satisfying strong consistency although seeking such a condition is of theoretical interest in model selection (in fact, many information criteria are to be satisfactory even if consistency is not achieved).

The definitions of empirical entropy and the number of parameters are different in each problem to be considered. In 1979, Hannan-Quinn proved that for AR $d_n = 2 \log \log n$ is the minimum order satisfying strong consistency (Hannan-Quinn proposition). However, the same $d_n = 2 \log \log n$ has been applied to other problems as well as AR. In fact, the proof of the Hannan-Quinn proposition essentially depends on the properties of the AR problem, which is clear from the original paper by Hannan-Quinn, and the Hannan-Quinn proposition was not proved for any other problem including the LR problem. On the contrary, without noticing such a matter, the information criterion HQ was applied to those problems.

Recently, the Hannan-Quinn proposition has been proved for estimating classification rules which has many applications such as Markov order estimation, data mining, pattern recognition (Suzuki, 2006).

This paper shows that the Hannan-Quinn proposition is true for estimating dependencies in LR, which seems to be of great significance. Otherwise, there would be no reason to use HQ in LR. Several authors suggested that $d_n = c \log \log n$ with some positive constant c would be enough (Rao-Wu, 1989). So, there has been evidence that the proposition is true although no formal proof appeared. This paper proves that such a c is any constant strictly greater than two.

In Section 2, we briefly overview how the Hannan-Quinn proposition was proved in AR. In Section 3, we derive the asymptotic error probability of model selection in LR

when information criteria are applied, which will be an important step to prove the main result. In Section 4, we give a proof of the Hannan-Quinn proposition for LR. Section 5 summarizes the results in this paper and gives a future problem.

Throughout the paper, we denote by $X(\Omega)$ the image $\{X(\omega)|\omega \in \Omega\}$ of a random variable $X : \Omega \to \mathbb{R}$, where Ω is the underlying sample space.

2 Auto Regression

Let $\{W_i\}_{i=-\infty}^{\infty}$ be a sequence of independent and identically distributed random variables with expectation zero and variance one, and let $\{X_i\}_{i=-\infty}^{\infty}$ be defined by

$$X_i = \sum_{j=1}^{k} \lambda_j X_{i-j} + W_i$$

and a nonnegative real sequence $\{\lambda_i\}_{i=1}^{k}$, where we assume the expectation of each X_i to be zero. Since $\{X_i\}$ is stationary, we obtain for $m \geq 0$, the following equation (Yule-Waker)

$$\gamma_m = \sum_{j=1}^{k} \lambda_j \gamma_{m-j} + \delta_{0m}\sigma_k^2 \ ,$$

where $\gamma_m := EX_iX_{i+m}$ does not depend on i. Using Cramer's formula, and from the values of $\{\gamma_m\}_{m=0}^{k}$, we obtain the values of $\lambda_0 := \sigma_k^2$ and $\{\lambda_m\}_{m=1}^{k}$ as a solution of the $(k+1) \times (k+1)$ linear equations:

$$\begin{bmatrix} -1 & \gamma_1 & \gamma_2 & \cdots & \gamma_k \\ 0 & \gamma_0 & \gamma_1 & \cdots & \gamma_{k-1} \\ 0 & \gamma_1 & \gamma_0 & \cdots & \gamma_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \gamma_{k-1} & \gamma_{k-2} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \sigma_k^2 \\ \lambda_{1,k} \\ \lambda_{2,k} \\ \vdots \\ \lambda_{k,k} \end{bmatrix} = \begin{bmatrix} -\gamma_0 \\ -\gamma_1 \\ -\gamma_2 \\ \vdots \\ -\gamma_k \end{bmatrix} .$$

Since the values of $\{\gamma_m\}_{m=0}^{k}$ are generally unknown, we need to estimate

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\hat{\gamma}_m := \hat{\gamma}_{-m} := \frac{1}{n}\sum_{i=1}^{n-m}(x_i - \bar{x})(x_{i+m} - \bar{x})$$

from the examples

$$x^n = (x_1, \cdots, x_n) \in X_1(\Omega) \times \cdots \times X_n(\Omega) \ .$$

Then, we obtain the Yule-Walker equation as follows:

$$\begin{bmatrix} -1 & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_k \\ 0 & \hat{\gamma}_0 & \hat{\gamma}_1 & \cdots & \hat{\gamma}_{k-1} \\ 0 & \hat{\gamma}_1 & \hat{\gamma}_0 & \cdots & \hat{\gamma}_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \hat{\gamma}_{k-1} & \hat{\gamma}_{k-2} & \cdots & \hat{\gamma}_0 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_k^2 \\ \hat{\lambda}_{1,k} \\ \hat{\lambda}_{2,k} \\ \vdots \\ \hat{\lambda}_{k,k} \end{bmatrix} = \begin{bmatrix} -\hat{\gamma}_0 \\ -\hat{\gamma}_1 \\ -\hat{\gamma}_2 \\ \vdots \\ -\hat{\gamma}_k \end{bmatrix} . \tag{2}$$

In particular, if the order k is unknown, we solve the above linear equation for each k to calculate the value of

$$L(x^n, k) = \frac{n}{2} \log \hat{\sigma}_k^2 + \frac{k}{2} d_n \ . \tag{3}$$

We estimate the true $k = k^*$ by the one $k = \hat{k}$ that minimizes (3). This process is called estimating the AR order. Then, we also obtain the solutions $\hat{\lambda}_{0,\hat{k}} := \hat{\sigma}_{\hat{k}}^2$ and $\{\lambda_{m,\hat{k}}\}_{m=1}^{\hat{k}}$ of (2) with $k = \hat{k}$.

In general,

$$\hat{\sigma}_k^2 = \{1 - \hat{\lambda}_{k,k}^2\} \hat{\sigma}_{k-1}^2 \ ,$$

thus for each $k = 1, 2, \cdots$, we have

$$\begin{aligned} & 2\{L(x^n, k) - L(x^n, k-1)\} \\ = \ & n \log \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2} + d_n \\ \leq \ & -n(1 - \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2}) + d_n \\ = \ & -n\hat{\lambda}_{k,k}^2 + d_n \ . \end{aligned} \tag{4}$$

As $n \to \infty$, for $k \leq k^*$, $\dfrac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2}$ almost surely converges to a value less than one. Thus, from (4), we have with probability one

$$L(x^n, 0) > L(x^n, 1) > \cdots > L(x^n, k^* - 1) > L(x^n, k^*) \ .$$

On the other hand, for $k \geq k^* + 1$, $\dfrac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2}$ almost surely converges to one. Hannan-Quinn(1979) proved from the law of iterated logarithms that

$$\frac{\hat{\lambda}_{k,k}^2}{2n^{-1} \log \log n} \leq 1$$

with probability one, and that for $d_n = 2c \log \log n$ $(c > 1)$,

$$L(x^n, k^*) < L(x^n, k^* + 1) < \cdots$$

with probability one.

3 Linear Regression

Let $X_1, \cdots, X_m$ be random variables such that there are no linear relations: any linear combination of $X_1, \cdots, X_m$ cannot be zero with probability one. Let $\epsilon \sim \mathcal{N}(0, \sigma^2)$ be a normal random variable with expectation zero and variance $\sigma^2 > 0$, and

$$Y := \sum_{j=1}^{p} \alpha_j X_j + \epsilon \ ,$$

where $\boldsymbol{\alpha} := [\alpha_1, \cdots, \alpha_p]^T \in \mathbb{R}^p$ $(0 \le p \le m)$. We assume that ϵ is independent of any linear combination of $X_1, \cdots, X_m$.

Suppose we do not know the values of order p and coefficients $\boldsymbol{\alpha}$, and that we are given independently emitted n examples

$$z^n := \{[y_i, x_{i,1}, \cdots, x_{i,m}]\}_{i=1}^n$$

with

$$y_i \in Y(\Omega), [x_{i,1}, \cdots, x_{i,m}] \in X_1(\Omega) \times \cdots \times X_m(\Omega) \ ,$$

where $\{[x_{1,j}, \cdots, x_{n,j}]\}_{j=1}^m$ are to be linearly independent. If we define

$$\boldsymbol{X}_p := \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{bmatrix}, \ \mathbf{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \ ,$$

we can write $\boldsymbol{y} = \boldsymbol{X}_p \boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Suppose that we estimate p by q $(0 \le q \le m)$. If we wish to minimize the quantity $\sum_{i=1}^n (y_i - \sum_{j=1}^q \hat{\alpha}_{jq} x_{ij})^2$ given the n examples, then $\hat{\boldsymbol{\alpha}}_q = [\hat{\alpha}_{1,q}, \cdots, \hat{\alpha}_{q,q}]^T := (\boldsymbol{X}_q^T \boldsymbol{X}_q)^{-1} \boldsymbol{X}_q^T \boldsymbol{y}$ is the exact solution (minimum square error estimation), where

$$\boldsymbol{X}_q := \begin{bmatrix} x_{1,1} & \dots & x_{1,q} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,q} \end{bmatrix}$$

3.1 Idempotent Matrices

Suppose $p \le q$. If we define $P_q := \boldsymbol{X}_q (\boldsymbol{X}_q^T \boldsymbol{X}_q)^{-1} \boldsymbol{X}_q^T$, we have

$$P_q^2 = P_q$$

and

$$(I - P_q)^2 = I - P_q \ ,$$

so that the square error is expressed by

$$\begin{aligned} S_q &:= \sum_{i=1}^n (y_i - \sum_{j=1}^q \hat{\alpha}_{j,q} x_{i,j})^2 \\ &= ||\boldsymbol{y} - \boldsymbol{X}_q \hat{\boldsymbol{\alpha}}_q||^2 \\ &= ||(I - P_q)\boldsymbol{y}||^2 \\ &= \boldsymbol{y}^T (I - P_q) \boldsymbol{y} \ . \end{aligned}$$

Similarly, if $q = p$, for $P_p := \boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T$ and $\hat{\boldsymbol{\alpha}}_p = [\hat{\alpha}_{1,p}, \cdots, \hat{\alpha}_{p,p}]^T := (\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{y}$, the square error is expressed by

$$S_p = \boldsymbol{y}^T(I - P_p)\boldsymbol{y} \ .$$

Thus, the difference between the square errors is

$$S_p - S_q = \boldsymbol{y}^T(I - P_q)\boldsymbol{y} - \boldsymbol{y}^T(I - P_q)\boldsymbol{y} = \boldsymbol{y}^T(P_q - P_p)\boldsymbol{y} \ . \tag{5}$$

On the other hand, we have

$$\begin{aligned} P_q^T &= (X_q^T)^T\{(X_q^TX_q)^{-1}\}^TX_q^T \\ &= X_q\{(X_q^TX_q)^T\}^{-1}X_q^T = P_q \end{aligned}$$

and $P_p^T = P_p$. From $P_qX_p = X_p$, $P_pX_p = X_p$, we obtain

$$P_qP_p = P_qX_p(X_p^TX_p)^{-1}X_p^T = X_p(X_p^TX_p)^{-1}X_p^T = P_p$$

and

$$P_pP_q = P_p^TP_q^T = (P_qP_p)^T = P_p^T = P_p \ .$$

Thus, not just for $P_p, I - P_p$ but also for $P_q - P_p$, the property

$$(P_q - P_p)^2 = P_q^2 - P_qP_p - P_pP_q + P_p^2 = P_q - P_p$$

holds. Such square matrices satisfying the property are called idempotent matrices (Chatterjee-Hadi, 1987).

In general, for idempotent matrix $P \in \mathbb{R}^{n\times n}$, the inner product $(Px, (I-P)x) = 0$ for any $x = Px + (I-P)x \in \mathbb{R}^n$, so that the eigenspaces are

1. $V_1 := \{Px | x \in \mathbb{R}^n\}$ with $\dim(V_1) = \mathrm{rank}(P)$, and
2. $V_0 := \{(I-P)x | x \in \mathbb{R}^n\}$ with $\dim(V_0) = n - \mathrm{rank}(P)$.

Since the eigenvalues are one and zero, the multiplicity of eigenvalue one is the same as the trace. Notice that for $(X_q^TX_q) = [y_{jk}]$ and $(X_q^TX_q)^{-1} = [z_{jk}]$,

$$trace(P_q) = trace(X_q(X_q^TX_q)^{-1}X_q^T) = \sum_{i=1}^n\sum_{j=1}^q\sum_{k=1}^q x_{ij}z_{jk}x_{ki} = \sum_{j=1}^q\sum_{k=1}^q y_{kj}z_{jk} = \sum_{k=1}^q 1 = q \ ,$$

and $trace(P_p) = p$, so that we have the following table.

P	trace(P)	$\dim(V_1)$	$\dim(V_0)$	rank(P)
P_p	p	p	$n-p$	p
$I-P_p$	$n-p$	$n-p$	p	$n-p$
P_q-P_p	$q-p$	$q-p$	$n-q+p$	$q-p$

3.2 Error probability in model selection

Proposition 1 *If $p < q$, $\dfrac{S_p - S_q}{S_p/n}$ asymptotically obeys the χ^2 distribution with freedom $q - p$.*

Proof: Given $\boldsymbol{X}_p$, we choose an orthogonal matrix $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n]$ of $I - P_p$ so that $U_1 =< \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{n-p} >$ and $U_0 =< \boldsymbol{u}_{n-p+1}, \cdots, \boldsymbol{u}_n >$ are the eigenspaces of eigenvalues one and zero, respectively. Notice that

$$(I - P_p)\boldsymbol{y} = \boldsymbol{y} - (X_p\boldsymbol{\alpha} + P_p\boldsymbol{\epsilon}) = \boldsymbol{\epsilon} - P_p\boldsymbol{\epsilon} = (I - P_p)\boldsymbol{\epsilon} \ . \tag{6}$$

For $j = 1, \cdots, n - p$, multiplying $\boldsymbol{u}_j^T$ in both hands from left, we get a normal random variable

$$z_j := \boldsymbol{u}_j^T \boldsymbol{y} = \boldsymbol{u}_j^T \boldsymbol{\epsilon} \ .$$

Since the expectation and variance of ϵ_i are zero and σ^2 (independent), and

$$\boldsymbol{u}_j^T \boldsymbol{u}_k = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} ,$$

we have $E[z_j] = 0$ and

$$E[z_j z_k] = E[\boldsymbol{u}_j^T \boldsymbol{\epsilon} \cdot \boldsymbol{u}_k^T \boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{u}_j^T \boldsymbol{u}_k = \begin{cases} \sigma^2, & j = k \\ 0, & j \neq k \end{cases} .$$

Thus, from the strong law of large numbers, with probability one as $n \to \infty$,

$$\frac{1}{n} S_p = \frac{1}{n} \sum_{j=1}^{n-p} z_j^2 \to \sigma^2 \ . \tag{7}$$

On the other hand, given $\boldsymbol{X}_q$, we choose an orthogonal matrix $V = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$ of $P_q - P_p$ so that $V_1 =< \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{q-p} >$ and $V_0 =< \boldsymbol{v}_{q-p+1}, \cdots, \boldsymbol{v}_n >$ are the eigenspaces of eigenvalues one and zero, respectively. Notice that from (6), we have

$$(P_q - P_p)\boldsymbol{y} = P_q(I - P_p)\boldsymbol{y} = P_q(I - P_p)\boldsymbol{\epsilon} = (P_q - P_p)\boldsymbol{\epsilon} \ .$$

For $j = 1, \cdots, q - p$, multiplying $\boldsymbol{v}_j$ in both hands from left, we get a normal random variable

$$r_j := \boldsymbol{v}_j^T \boldsymbol{y} = \boldsymbol{v}_j^T \boldsymbol{\epsilon} \ .$$

Since the expectation and variance of ϵ_i are zero and σ^2 (independent), and

$$\boldsymbol{v}_j^T \boldsymbol{v}_k = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} ,$$

we have $E[r_j] = 0$ and

$$E[r_j r_k] = E[\boldsymbol{v}_j^T \boldsymbol{\epsilon} \cdot \boldsymbol{v}_k^T \boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{v}_j^T \boldsymbol{v}_k^T = \begin{cases} \sigma^2, & j = k \\ 0, & j \neq k \end{cases} .$$

Hence, as $n \to \infty$,

$$\frac{S_p - S_q}{\sigma^2} = \sum_{j=1}^{q-p} \frac{r_j^2}{\sigma^2} \sim \chi_q^2 \tag{8}$$

where the fact that the square sum of $q-p$ independent random variables with the standard normal distribution obeys the χ^2 distribution of freedom $q-p$ has been applied. Equations (7)(8) imply Proposition 1.

(Q. E. D.)

In the sequel, for $\pi \subseteq \{1, \cdots, m\}$, we write the square error of $\{X_j\}_{j\in\pi}$ and Y by $S(\pi)$, and put

$$L(z^n, \pi) := n \log S(\pi) + \frac{k(\pi)}{2} d_n$$

and $k(\pi) = |\pi|$, given $z^n = \{[y_i, x_{i,1}, \cdots, x_{i,m}]\}_{i=1}^n$. Let $\pi_* \subseteq \{1, \cdots, m\}$ be the true π.

Theorem 1 For $\pi \supset \pi_*$, the probability of $L(z^n, \pi) < L(z^n, \pi_*)$ is

$$\int_{n\{1-\exp[-\frac{k(\pi)-k(\pi_*)}{2n} d_n]\}}^{\infty} f_{k(\pi)-k(\pi_*)}(x) dx \ ,$$

where f_l is the probability density function of the χ^2 distribution of freedom l.

Proof: Notice that

$$\begin{aligned} & 2\{L(z^n, \pi) - L(z^n, \pi_*)\} \\ = \ & 2n \log \frac{S(\pi)}{S(\pi_*)} + \{k(\pi) - k(\pi_*)\} d_n \\ = \ & 2n \log(1 - \frac{S(\pi_*) - S(\pi)}{S(\pi_*)}) + \{k(\pi) - k(\pi_*)\} d_n \ , \end{aligned} \tag{9}$$

so that

$$L(z^n, \pi) < L(z^n, \pi_*) \iff \frac{S(\pi_*) - S(\pi)}{S(\pi_*)/n} > n\{1 - \exp[-\frac{k(\pi) - k(\pi_*)}{2n} d_n]\} \ . \tag{10}$$

From Proposition 2, we obtain Theorem 1.

(Q. E. D.)

Hereafter, we do not assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ but that ϵ_i is an independently identically distributed random variable with expectation zero and variance σ^2.

Theorem 2 For $\pi \not\supseteq \pi_*$, $L(x^n, \pi) > L(x^n, \pi_*)$ with probability one as $n \to \infty$.

Proof: Suppose $q < p$. Given $\boldsymbol{X}_p$, we choose an orthogonal matrix $W := [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n]$ of $P_p - P_q$ so that $W_1 =< \boldsymbol{w}_1, \cdots, \boldsymbol{w}_{p-q} >$ and $W_0 =< \boldsymbol{w}_{p-q+1}, \cdots, \boldsymbol{w}_n >$ are the eigenspaces

of eigenvalue one and zero, respectively. Since $\{\hat{\alpha}_{j,p}\}_{j=1}^{p}$ are strongly consistent estimators (Lai-Robbins-Wei, 1978), we have for $j = 1, \cdots p-q$ with probability one as $n \to \infty$

$$\begin{aligned} s_j &:= \sum_{i=1}^{n} w_{i,j} y_i = \sum_{i=1}^{n} w_{ij} \{ \sum_{k=1}^{p} x_{ik} \hat{\alpha}_{k,p} + y_i - \sum_{k=1}^{p} x_{ik} \hat{\alpha}_{k,p} \} \\ &\to \sum_{i=1}^{n} w_{ij} (\sum_{k=1}^{p} x_{ik} \alpha_k + \epsilon_i) \\ &\to \sum_{i=1}^{n} w_{ij} (\sum_{k=q+1}^{p} x_{ik} \alpha_k + \epsilon_i) \ , \end{aligned}$$

where $\boldsymbol{w}_j := [w_{1,j}, \cdots, w_{n,j}]^T$. Since ϵ and $\sum_{k=q+1}^{p} \alpha_k X_k$ are independent, we have for $j = 1, \cdots, p-q$ with probability one as $n \to \infty$

$$\begin{aligned} \frac{1}{n} s_j^2 &\to (\sum_{i=1}^{n} w_{ij} \sum_{k=q+1}^{p} \frac{x_{ik}\alpha_k}{\sqrt{n}})^2 + \frac{\sigma^2}{n} \\ &\to (\sum_{i=1}^{n} w_{ij} \sum_{k=q+1}^{p} \frac{x_{ik}\alpha_k}{\sqrt{n}})^2 \end{aligned}$$

and

$$\boldsymbol{x}^n := (\sum_{k=q+1}^{p} \frac{x_{1,k}\alpha_k}{\sqrt{n}}, \cdots, \sum_{k=q+1}^{p} \frac{x_{n,k}\alpha_k}{\sqrt{n}})$$

has a positive constant square norm $||\boldsymbol{x}^\infty||^2$ as $n \to \infty$ unless $\sum_{k=q+1}^{p} \alpha_k X_k = 0$ with probability one, which contradicts our assumption. Since $\boldsymbol{x}^n$ is not orthogonal to the space $< \boldsymbol{w}_1, \cdots, \boldsymbol{w}_{p-q} >$ and $||\boldsymbol{x}^\infty||^2 > 0$, from (5),

$$\frac{1}{n}(S_q - S_p) \to \lim_{n\to\infty} \sum_{j=q+1}^{p} (\boldsymbol{w}_j^T \boldsymbol{x}^n)^2 > 0 \ , \tag{11}$$

which implies the theorem when $\pi \subset \pi_*$. Suppose $\pi \not\subset \pi_*$. In the same way, if we notice that (11) is true even for $q = |\pi \cap \pi_*|$, so that

$$\lim_{n\to\infty} \frac{1}{n} \{ S(\pi \cap \pi_*) - S(\pi_*) \} > 0 \ . \tag{12}$$

Furthermore, if we replace π_* by $\pi \cap \pi_*$, from a similar discussion as in Theorem 1, we have

$$\lim_{n\to\infty} \frac{1}{n} \{ S(\pi) - S(\pi \cap \pi_*) \} = 0 \ . \tag{13}$$

The statements (12)(13) imply the theorem.

(Q. E. D.)

4 Proof of the Hannan-Quinn Proposition

Proposition 2 *If $q > p$, with probability one,*

$$\frac{S_p - S_q}{S_p} \le (q-p)\log\log n \tag{14}$$

Proof: The notation is similar to Proposition 2, and let $p+1 \le j \le q$. For $Z_i := \dfrac{\sqrt{n}v_{i,j}\epsilon_i}{\sigma}$ with $\boldsymbol{v}_j = [v_{1,j}, \cdots, v_{n,j}]^T$, we have $\displaystyle\sum_{i=1}^n Z_i = \frac{\sqrt{n}r_j}{\sigma}$ with expectation zero and variance σ^2, and $E[\sum_{i=1}^n Z_i] = 0$, $E[(\sum_{i=1}^n Z_i)^2] = n$. Since Z_i is independently identically distributed. $E[Z_i] = 0$, $E[Z_i^2] = 1$. From the law of iterated logarithms (Stout 1974), we have

$$\frac{\sum_{i=1}^n Z_i}{\sqrt{n\log\log n}} = \frac{\sqrt{n}\boldsymbol{v}_j^T\boldsymbol{\epsilon}/\sigma}{\sqrt{n\log\log n}} \le 1\ ,$$

namely,

$$\frac{r_j}{\sigma} \le \sqrt{\log\log n}$$

with probability one, which means

$$\frac{S_p - S_q}{S_p/n} \le (q-p)\log\log n$$

with probability one.

(Q. E. D.)

Theorem 3 For $d_n := 2c\log\log n$ $(c > 1)$, $L(z^n, \pi) > L(z^n, \pi_*)$ with probability one.

Proof: From Theorem 2, the error for $\pi_* \not\subseteq \pi$ is almost surely zero as long as $\dfrac{d_n}{n} \to 0$ $(n \to \infty)$, so that we only need to consider the case $\pi_* \subset \pi$. However, $d_n = 2c\log\log n$ with $c > 1$ implies the both sides of

$$\frac{1}{2}\{k(\pi)-k(\pi_*)\}d_n - \frac{1}{4n}[\{k(\pi)-k(\pi_*)\}d_n]^2 \le n[1-\exp\{-\frac{k(\pi)-k(\pi_*)}{2n}d_n\}] \le \frac{1}{2}\{k(\pi)-k(\pi_*)\}d_n$$

(see (10)) are at least $(q-p)\log\log n$ with $p = k(\pi_*)$ and $q = k(\pi)$ for large n (Proposition 2), which implies Theorem 3.

(Q. E. D.)

5 Conclusion

We proved that the Hannan-Quinn proposition is true for linear regression as well as for auto regression (Hannan-Quinn, 1979) and for classification (Suzuki, 2006): the minimum rate of d_n satisfying strong consistency is $(2+\epsilon)\log\log n$ for arbitrary $\epsilon > 0$.

The future problems contain finding strong consistency conditions that are good for all the cases including linear regression, auto regression, and classification. Making clear why the same $d_n = 2\log\log n$ is the crucial rate for those problems would be the first step to solve the problem.

References

[1] Akaike, H. (1974): "A New Look at the Statistical Model Identification," I.E.E.E. Transactions on Automatic Control, AC 19, 716-723

[2] Schwarz, G. (1978): "Estimating the Dimension of a Model," Annals of Statistics, 6, 461-464.

[3] Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," Journal of the Royal Statistical Society, B, 41, 190-195.

[4] J. Suzuki (2006): On Strong Consistency of Model Selection in Classification. IEEE Transactions on Information Theory 52(11): 4767-4774

[5] Rao, C.R., Wu, Y., (1989): A strongly consistent procedure for model selection in a regression problem. Biometrika 76, 369-374

[6] Chatterjee, S. and Hadi, A. S. (1988), Sensitivity Analysis In Linear Regression, New York: John Wiley & Sons.

[7] T.L. Lai, H. Robbins, C.Z. Wei (1978): "Strong consistency of least squares estimates in multiple regression". Proceedings of the National Academy of Sciences USA 75 (7).

[8] Stout, W. (1974). Almost Sure Convergence. New York: Academic Press.